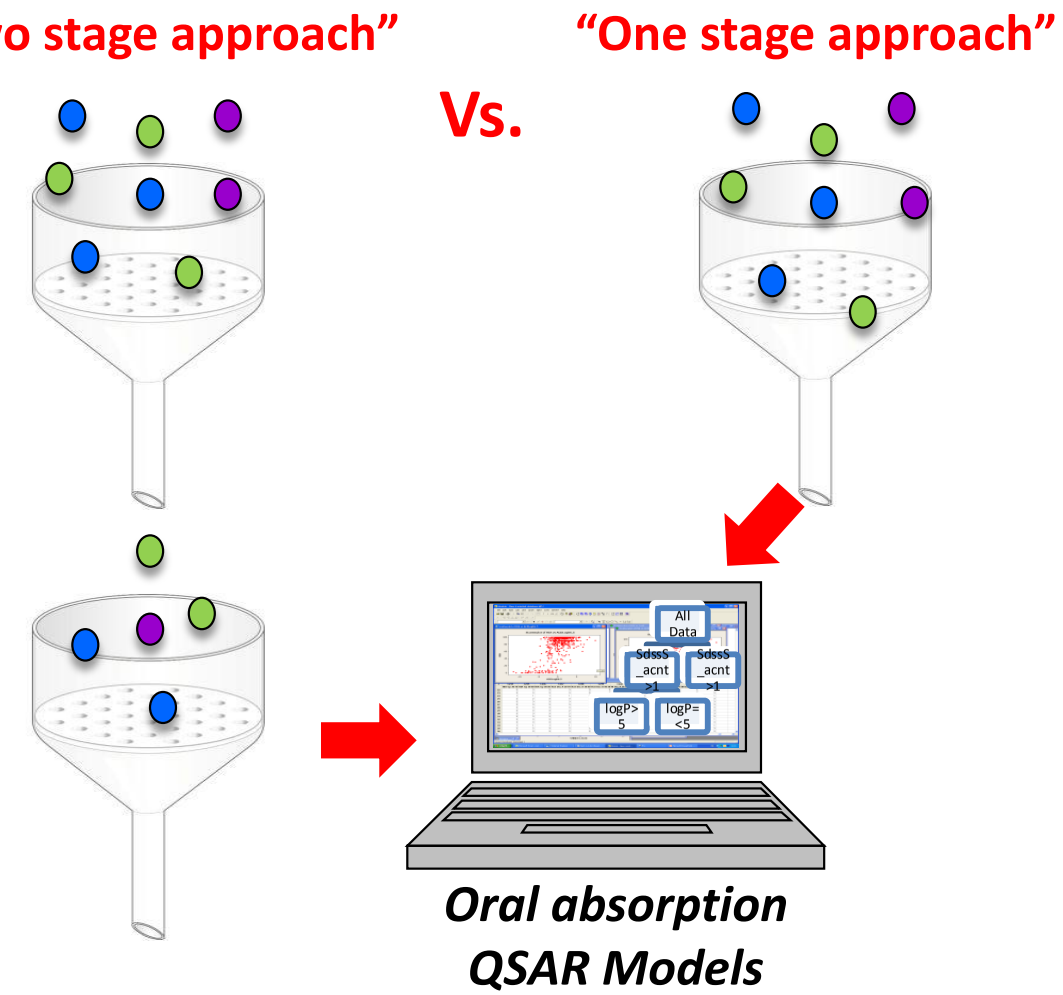
Pre-processing feature selection for improved C&RT models for oral absorption

Article (Accepted Version)


Pre-processing feature selection
"Two stage approach"
Vs.
No pre-processing
"One stage approach"
All Data
SdssS _acnt >1
SdssS _acnt >1
logP> 5
logP= <5
Oral absorption
QSAR Models

Pre-processing feature selection for improved C&RT models for oral absorption

Danielle Newby[a], Alex. A. Freitas[b], Taravat Ghafourian[a,c]*

[a]*Medway School of Pharmacy, Universities of Kent and Greenwich, Chatham, Kent, ME4 4TB, UK*

[b]*School of Computing, University of Kent, Canterbury, Kent, CT2 7NF, UK*

[c] *Drug Applied Research Centre and Faculty of Pharmacy, Tabriz University of Medical Sciences, Tabriz, Iran*

*** Corresponding Author,** Email: T.ghafourian@kent.ac.uk; Tel +44(0)1634 202952; Fax +44 (0)1634 883927

Abstract

There are currently thousands of molecular descriptors that can be calculated to represent a chemical compound. Utilising all molecular descriptors in Quantitative Structure-Activity Relationships (QSAR) modelling can result in overfitting, decreased interpretability and thus reduced model performance. Feature selection methods can overcome some of these problems by drastically reducing the number of molecular descriptors and selecting the molecular descriptors relevant to the property being predicted. In particular, decision trees such as C&RT, although they have an embedded feature selection algorithm, can be inadequate since further down the tree there are fewer compounds available for descriptor selection and therefore descriptors may be selected which are not optimal. In this work we compare two broad approaches for feature selection: (1) a "two-stage" feature selection procedure, where a pre-processing feature selection method selects a subset of descriptors, and then classification and regression trees (C&RT) selects descriptors from this subset to build a decision tree; (2) a "one-stage" approach where C&RT is used as the only feature selection technique. These methods were applied in order to improve prediction accuracy of QSAR models for oral absorption. Additionally, this work utilises misclassification costs in model building to overcome the problem of the biased oral absorption datasets with more highly-absorbed than poorly-absorbed compounds. In most cases the two stage feature selection with pre-processing approach had higher model accuracy compared with the one stage approach. Using the top 20 molecular descriptors from random forest predictor importance method gave the most accurate C&RT classification model. The molecular descriptors selected by the five filter feature selection methods have been compared in relation to oral absorption. In conclusion, the use of filter pre-processing feature selection methods and misclassification costs produce models with better interpretability and predictability for the prediction of oral absorption.

Keywords:

Oral absorption, intestinal absorption, in silico, classification, feature selection, QSAR

1. Introduction

The cost of bringing a drug to the market keeps on increasing [1, 2]. The expense is likely to rise further with higher costs of everything from consumables to clinical studies and also tighter regulations governing acceptance of oral drugs on the market [3]. Although there has been a successful effort to reduce compound attrition rates by incorporating pharmacokinetic (PK) assays in a high throughput manner earlier in drug discovery, compounds are now failing for other reasons as well as poor PK such as efficacy and toxicity [4]. There is specific interest in predicting the intestinal absorption of new chemical entities (NCEs) as the oral route is the dominant route of drug delivery due to ease of administration and patient acceptance [5, 6]. *In silico* modelling of intestinal absorption using QSAR (Quantitative Structure-Activity Relationships) can be used as a cost effective strategy to remove unsuitable compounds based on physicochemical properties and chemical structure alone. Moreover, *in silico* modelling can be used in tandem with high throughput assays in drug discovery and act as a guide to select appropriate assays that will help understand the mechanistic absorption properties of compounds [7].

QSAR involves the mathematical relationships between a molecular structure and biological activity. However this relationship cannot be determined directly, therefore molecular descriptors that describe the chemical structure are calculated to derive relationships between the molecular descriptors and activity. Molecular descriptors are numerical representations of the chemical structure. Molecular descriptors can be classed as 0, 1, 2, 3 and 4D groups [8]. Simple 0D descriptors are counts of atom and bonds in structure such as molecular weight and number of hydrogen atoms in a molecule. Molecular descriptors that count structural fragments, atomic properties or fingerprints are classed as 1D. Examples of 1D are number of hydrogen bond donors or acceptors. Topological descriptors based on the 2D structure of the molecule are predicted using graph theory, vectors and indices, and examples include the kappa shape, chi connectivity indices [9] and topological polar surface area [10]. More complicated molecular descriptors such as 3D and 4D require the 3D coordinates of the structure. 3D descriptors are geometric descriptors and there are two types based on the internal or external orientation properties of the molecule. Good examples of 3D descriptors are energies relating to the orbitals of the atoms in the compound such as the lowest unoccupied molecular orbital (LUMO) and the highest occupied molecular orbital (HOMO) energies. These molecular descriptors are derived from quantum chemistry theories and relate

to the reactivity of the compound. Finally 4D descriptors are based on the 3D structure but take into account the different flexibilities of the structure [8].

In order to produce a model that is robust and high in predictive power, a wide choice of molecular descriptors is very important. Identifying the relevant descriptors correlating with intestinal absorption can be carried out using statistical feature selection methods although, additionally, educated assumptions can be made about physiological and physicochemical factors that influence the process of oral absorption to choose the useful descriptors [11].

Feature selection is used frequently in QSAR and data mining to selectively minimise the number of independent variables (molecular descriptors) used to accurately describe the dependent variable – i.e., absorption [12]. Feature selection is important for numerous reasons. Firstly, fewer molecular descriptors increase interpretability and understanding of resulting models [13, 14]. Secondly feature selection can provide improved model performance for the prediction of new compounds [15, 16]. Finally, it can reduce the risk of overfitting from noisy redundant molecular descriptors [17].

Feature selection can be split into two broad categories: data pre-processing or embedded methods. Data pre-processing feature selection involves reduction of the number of molecular descriptors before model building, unlike embedded methods that incorporate the feature selection into the training and building of the model [17, 18]. Data pre-processing techniques can be further split into filter and wrapper techniques. Filter techniques usually involve calculating a relative score of the molecular descriptors and ranking them in order of best score, and the descriptors that are at the top of the list are then used as input for classification. Examples of these are chi square and information gain. Wrapper techniques consider a number of subsets of molecular descriptors, evaluate each of these based on the predictive performance of a classification model built from that descriptor subset and eventually select the descriptor subset with the best predictive performance [19]. In comparison of filter and wrapper methods, there are advantages and disadvantages. The choice of method depends on many things such as interpretability, predictability and computational cost.

Filter methods offer a fast and simple way to select important descriptors. In addition, because they are independent of the classification algorithm, the score for each descriptor only needs to be calculated once, and the selected descriptors can be used as input for a variety of classification algorithms. A disadvantage of univariate filter methods is they fail to account for interactions between independent variables as most measure the correlation

between the dependent variable and each independent variable separately. This can be overcome by multivariate methods which take into account independent variable interactions.

Wrapper techniques on the other hand are usually more computationally expensive, but unlike many univariate filter techniques, they take into account independent variable interactions [17, 18]. In addition, hybrid filter and wrapper methods have also been developed as successful feature selection techniques [20]

Most oral absorption models in the literature have utilised feature selection methods either in pre-processing or in model development. There are many types of research in the literature that focus on different issues of oral absorption modelling; e.g. those that focus on obtaining a high predictive model with the feature selection not as the primary focus, but just as a part of the modelling process [21]; and those that compare different feature selection techniques and compare the molecular descriptors chosen by the different techniques [11, 20]. However, an underlying problem of oral absorption models in the literature is that they were developed using current oral absorption datasets in the literature which are highly biased towards the prediction of highly-absorbed compounds [21-24]. This is due to availability of more data on marketed drugs which are mostly highly-absorbed in contrast with data on compound and drug candidates that never made into the market and failed during drug discovery. The models in this case may predict high absorption rate for poorly-absorbed compounds, i.e. false positives. This is not an ideal scenario as in drug discovery more compounds are now poorly-absorbed due to higher lipophilicity and poor aqueous solubility of current drug candidates [25, 26].

Two methods have been studied previously to overcome the problem of biased oral datasets that show the effect of data distribution in the training sets for regression and classification. Firstly under-sampling the majority class, highly-absorbed compounds, to create a balanced training set with the same number of poorly and highly-absorbed compounds [27]. The second technique utilises the whole biased dataset but applies misclassification costs to reduce false positives [28]. The use of higher misclassification costs for model development should improve the predictive power of the model built with the molecular descriptor subsets chosen by appropriate feature selection methods.

This work investigates five pre-processing filter feature selection techniques for selecting subsets of molecular descriptors. The comparison of these different feature selection techniques is anticipated to give an idea of the relative abilities of the different techniques

based on their prediction ability on the validation set. Furthermore, we compare two broad approaches for feature selection: (1) a "two-stage" feature selection procedure, where in the first stage a pre-processing feature selection method selects a subset of descriptors, and in the second stage classification and regression trees (C&RT), which is itself an embedded feature selection method, selects a subset of the descriptors selected by the filter technique to build a decision tree; (2) a "one-stage" approach where C&RT is used as the only feature selection technique, without using data pre-processing feature selection methods. A comparison between these two approaches could indicate the usefulness of pre-processing feature selection for C&RT analysis. Additionally, this work utilises misclassification costs in model building to overcome the problem of biased datasets. This work offers an investigation of feature selection techniques which reduces the number of molecular descriptors, increasing interpretability of resulting models and combined with this the use of misclassification costs in model development to increase model predictability when analyzing a biased dataset. Therefore this work offers a novel combination of pre-processing feature selection combined with misclassification costs to develop models for biased oral absorption datasets.

2. Methods and Materials

2.1 Dataset and Misclassification Costs

The published dataset of Hou et al[21] containing %HIA (Percent Human Intestinal Absorption) data for 645 drugs and drug-like compounds was utilised for development and optimisation of models. An additional set of data was collated from literature to serve as the external validation set. The %HIA values and references for the external validation set can be found in the **Supporting Information**.

All the compounds in Hou et al's data set were sorted by ascending %HIA values and then by logP values. The %HIA ascending values were put into groups of six then 5/6th of these compounds were placed randomly in the training set and the remaining into the parameter optimisation set (internal test set). The training set was used to train the model in C&RT; the parameter optimisation set was used to obtain the best parameters for the models. In addition, the external validation set was used to show the predictive ability of the models created with an unseen validation set. All compound sets had similar data distributions of highly and poorly-absorbed compounds to create a fairer more controlled validation of the models. The

exact number of compounds in the training, parameter optimisation and validation set are shown in **Table 1**.

Table 1. Numbers of Compounds for training, parameter optimisation and validation sets

Data Set	Number of compounds (N)
Training set	534
Parameter optimisation set	107
Validation set	48

As stated previously the data set is highly skewed with many more highly-absorbed than poorly-absorbed compounds. Therefore any model generated using this biased dataset will be better at predicting highly-absorbed than poorly-absorbed compounds and there will be more misclassified poorly-absorbed compounds (false positives). To overcome this problem applying a higher misclassification cost to the poorly-absorbed misclassification (false positive) will reduce the number of false positives and increase overall prediction accuracy. In a previous investigation it was shown that applying misclassification costs to the prediction of poorly-absorbed compounds improved the predictive power especially for poorly-absorbed compounds by overcoming the distribution bias of the dataset [28]. In this work, in order to assign an objective number for the overall misclassification cost, we have used the class distribution of the highly and poorly-absorbed compounds. Therefore we have used a misclassification factor of four to one, for low and high classes, respectively.

2.2 Molecular descriptors

A variety of different software packages were used to compute molecular descriptors; they include TSAR 3D v3.3 (Accelrys Inc.), MDL QSAR (Accelrys Inc.), MOE (Chemical Computing Group Inc.) v2010.10 and Advanced Chemistry Development ACD Labs/ LogD Suit v12. A total of 204 descriptors were initially used in this study before applying feature selection methods.

2.3 Classification and Regression Trees (C&RT)

Classification of the compounds using C&RT analysis was carried out using STATISTICA v11 (StatSoft Ltd). Compounds were placed into categorical classes of ‘high’ or ‘low’ according to the observed %HIA value in the dataset. The threshold for the classes was 50%; therefore any compounds with %HIA $\geq$ 50% was assigned to the ‘High’ class and any compound with a %HIA less than 50% was assigned to the ‘Low’ class.

C&RT analysis is a statistical technique that uses decision trees to solve regression and classification problems [29]. For this work, the dependent variable (HIA Class) was categorical and classification trees were produced which classed compounds either 'high' or 'low' absorption. For this work the stopping factors were minimum number of compounds for splitting at 30 based on preliminary experiments. This enables pruning of the tree and prevents over-fitting of the decision tree [29, 30].

For this work, HIA Class was set as the dependent categorical variable and either all 203 molecular descriptors or a subset of these selected by various feature selection methods were selected as continuous independent variables. The analyses also included one categorical independent variable, N+ group, the indicator variable for presence or absence of quaternary ammonium. If there were any trees with only one compound in the terminal nodes, manual pruning was carried out to prevent this final split so that no terminal nodes contained only one compound. All other settings used were default setting defined by the software.

It must be noted that C&RT performs embedded feature selection; therefore in this work we are also investigating the use of feature selection methods in a pre-processing phase, before inputting the descriptor subset into C&RT. By carrying out data pre-processing feature selection the methods can avoid C&RT's drawback of 'data fragmentation'. In other words, as the decision tree is built and compounds split into smaller nodes there are fewer compounds to split; therefore, the selection of descriptors in that local node becomes less statistically reliable. **Figure 1** shows the work flow of this investigation and how the pre-processing feature selection selects molecular descriptors as input for C&RT analysis compared to the embedded C&RT approach.

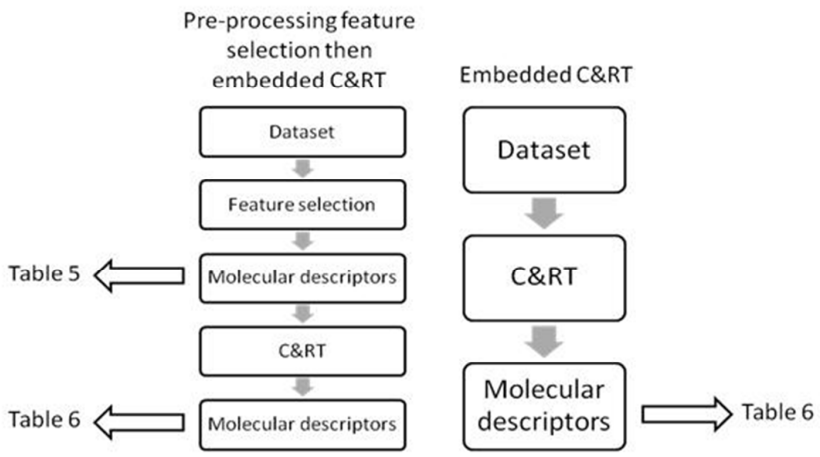


Figure 1: Workflow for molecular descriptor generation for pre-processing feature selection and embedded C&RT analysis

2.4 Missing values

Missing values for molecular descriptors can be a problem when building QSAR models. Depending on the software, procedures used to overcome the problems of missing values will vary [20, 31]. For example, with general C&RT analysis in STATISTICA any compounds with missing values for certain molecular descriptors will be removed at the tree root. This means that there are fewer compounds used to build the C&RT models and the possibility of reducing the chemical coverage of the resulting QSAR model. In comparison, interactive trees will remove chemical compounds from the decision tree on a case by case basis, so only when that particular molecular descriptor is picked in the C&RT analysis will the chemical compounds be removed. Missing molecular descriptors values for compounds can identify patterns relating to certain functional groups or structural features that give rise to the missing values. In this work it was noted that compounds that contained a permanent quaternary ammonium ion had more missing descriptor values than other compounds in the dataset. Therefore, an indicator variable that described the permanent positive nitrogen (YES/NO) was calculated. Molecular descriptors that are difficult to compute and result in missing values may not be suitable to be used in resulting models as the molecular descriptors may not be able to be calculated for new compounds, leading to poor performance of the model for classification of these compounds. Therefore, we removed all molecular descriptors that had 10 or more missing values based on preliminary work, and therefore had a final number of 204 descriptors available for feature selection techniques.

2.5 Feature Selection

We used feature selection methods in pre-processing step to reduce the number of molecular descriptors to a smaller subset that accurately describes the dependent variable, in this case HIA Class. The software used for feature selection was STATISTICA v11 and WEKA v 3.6 [32]. The feature selection techniques to select molecular descriptors for the classification models of oral absorption are shown in **Table 2**. The descriptors selected by the feature selection techniques in **Table 2** were used as input by C&RT which then performed further (embedded) feature selection (**Figure 1**).

Table 2. Pre –processing feature selection methods utilised in this work

	Feature selection method	Acronym used in this paper	Software used
1	Predictor importance using random forest	RF	STATISTICA
2	Predictor importance using random forest with higher misclassification costs for false positives	RF (MC)	STATISTICA
3	Chi-square	CS	STATISTICA
4	Information gain ratio	IGR	WEKA
5	Greedy stepwise	GRD	WEKA
6	Genetic search	GEN	WEKA

It is also important to define which parts of dataset were used for the different feature selection techniques. The training set is used by all methods; however, for the filter methods CS, IGR, GRD and GEN the parameter optimisation set was combined with the training set to carry out feature selection using these techniques. For random forest and C&RT (embedded feature selection) the training set was used to train the model and separately the parameter optimisation set was used to obtain optimal parameters for the method (**Figure 2**).

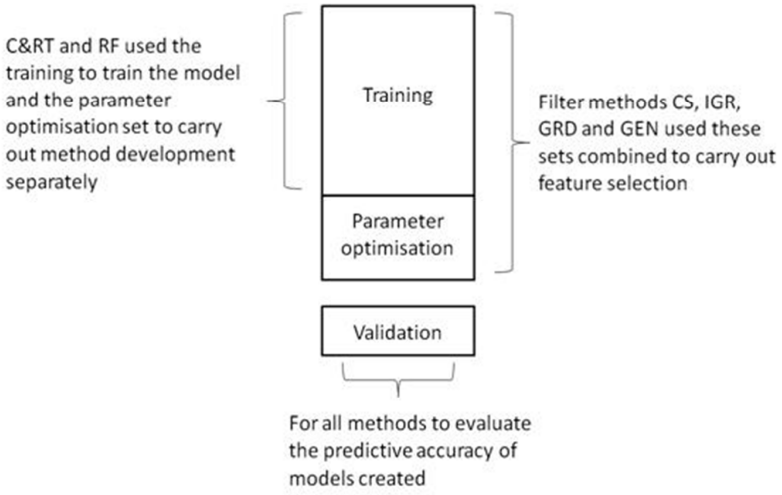


Figure 2. Compound sets were used for pre-processing and embedded feature selection.

In this work for methods RF, CS, IGR the top 20 molecular descriptors were selected based on the highest values of the descriptor scoring function. Other numbers of selected molecular descriptors were tried; however, based on the C&RT analysis results on the parameter optimisation set, the top 20 descriptors gave the highest classification accuracy and was selected.

2.5.1 Predictor importance ranking using Random forest (RF)

Random forest generates a set of decision trees based on random subsets of compounds and descriptors in the training set. The ensemble of decision trees vote based on the individual tree results and then the majority vote for a particular compound determines the classification of that compound [33, 34].

This method was carried out for the training set and the parameters of the analysis were optimised using the parameter optimisation set. The top 20 descriptors based on a ranking function called predictor importance in STATISTICA were obtained from the selected model. In STATISTICA software, for every molecular descriptor, the drop in each node impurity (delta) is summed for all nodes in the trees and expressed relative to the largest sum – i.e. the most significant descriptor. The delta is calculated for every descriptor (even if not used in the node for the splitting of the tree) and summed for every node and tree produced in the forest. The larger the delta the more significant the molecular descriptor is. The final summed delta value for every descriptor is normalized against the most important molecular descriptor and therefore expressed relative to the molecular descriptors with the largest delta. This means that important molecular descriptors that may not have been picked to be in the trees may still appear in the final predictor importance table.

Optimization of the random forest method was carried out based on the plot of misclassification error on the parameter optimisation set vs. the number of trees. The misclassification rate is the number of misclassified compounds divided by the total number of compounds. The lower the misclassification rate for the parameter optimisation set, the better the model. Based on the misclassification rate, the optimum number of trees was selected and used to repeat the analysis again with the new optimized value. The maximum number of levels for each tree was set to three. The software default value of eight was used for the number of molecular descriptors used in each tree. For random forest there was an option to apply misclassification costs, therefore two sets of molecular descriptors were selected using this technique: a descriptor set selected using equal misclassification costs (RF) and a descriptor set selected using a misclassification cost ratio of 4:1 for false positives: false negatives (RF (MC)).

2.5.2 Chi Square (CS)

In STATISTICA the CS function can be calculated and molecular descriptors ranked accordingly. CS is a statistical measure of the association (or dependence) between two

categorical variables [35]. The greater the CS value, the more statistically significant the molecular descriptor is in relation to the %HIA class, therefore allowing the most statistically important molecular descriptors to be ranked. The main drawback of using CS as well as many other filter techniques is that it is a univariate feature selection method; therefore it does not take into account interactions between the molecular descriptors. This could be a potential issue in relation to intestinal absorption, where there are many interlinking factors influencing absorption with many molecular descriptors describing them [6, 36]. CS is an association measure for categorical descriptors, therefore there may be problems when continuous variables are used that contain a large spread of numerical values, since the conversion of numerical variables into categorical ones (required for the use of the chi square measure) may lose relevant information. The software default number of bins (ten) was used for chi square discretizing of the molecular descriptors.

2.5.3 Information gain ratio (IGR)

Information gain ratio is a normalised function of the information gain feature selection method developed by Quinlan[37] as part of the ID3 (Iterative Dichotomiser) decision tree algorithm. This feature selection method is used to split the decision tree into nodes and identify molecular descriptors that are the best for the individual splits [37]. Information gain works to minimise the information needed to classify compounds into resulting nodes. It is the difference between the original information (before the data is spilt) and the new information produced after using the molecular descriptor to split the training set data. This difference is the gain of information achieved by using a specific molecular descriptor, therefore the molecular descriptor with the highest gain is the one used for the split [14]. Information gain ratio was first described by Quinlan[38] in the context of the C4.5 algorithm, which superseded ID3. Information gain ratio overcomes the bias towards selecting those molecular descriptors with many numerical values by normalising the information gain. The higher the ratio value the better the molecular descriptor for the split. This feature selection technique was carried out using WEKA 3.6.

2.5.4 Greedy Stepwise (GRD)

The previous feature selection methods are based on ranking the molecular descriptors based on a certain criteria and do not take into account the interactions between the molecular descriptors. Therefore two additional feature selection methods were used that utilise a search method which takes molecular descriptor interaction into account as well as the correlation with HIA class. These methods seek to maximise the correlation between HIA and the

molecular descriptors being tested, and minimise correlations between the molecular descriptors.

The first of these methods is greedy stepwise, which is a forward stepwise feature selection method [39]. This is a local search method that firstly considers all the molecular descriptors and picks the best one – i.e., the one that correlates with HIA class. It then starts again with all the remaining molecular descriptors, and picks the best molecular descriptor that pairs with the previously selected molecular descriptor in relation to HIA class. The iterations carry on until a local maximum is reached. Due to the nature of this technique only a local search can be carried out based on the molecular descriptor(s) selected in all the previous iterations, therefore the potential for a global search of all the different possible subsets is limited, and promising regions of molecular descriptor space can be missed [15]. To guide the greedy search in the feature selection process, in the WEKA software an evaluator is used. The evaluator function used was correlation-based feature selection subset evaluator (CfssubsetEval). This evaluator not only aims to maximise the correlation between the best molecular descriptors and HIA class, but also to minimise the correlation or redundancy between the descriptors for the search subsets generated.

2.5.5 Genetic Search (GEN)

GEN is a filter (rather than wrapper) version of the genetic algorithm [40]. Genetic algorithm (GA) was first created by Holland[41], although the concept of genetic algorithm was being researched before this. Now termed generally as an evolutionary algorithm, GA mimics the process of natural evolution. An initial population is created containing random candidate solutions. In the context of this work, a candidate solution is a molecular descriptor subset. Each candidate solution is evaluated in terms of its fitness (quality), and candidate solutions are then selected to be reproduced and to undergo modifications with a probability proportional to their fitness values. The process of selecting "parent" candidate solutions based on fitness and producing "offspring" solutions that are based on the parents is iteratively performed for a number of iterations, so that the population of candidate solutions gradually evolves towards better and better candidate solutions. [41]. In this work we have utilised the genetic search feature selection method using WEKA software [42]. This method carries out a global search in the 'molecular descriptor space' to find the best subset of molecular descriptors relating to HIA class, guided by a subset evaluator that generates a numerical value of 'fitness' (quality) of any given feature subset. Like with the greedy search technique, the evaluation function used for the genetic search method was 'CfssubsetEval'.

2.6 Statistical significance of the models

Specificity (SP), sensitivity (SE), cost normalized misclassification index (CNMI), and SP × SE were used to show the predictive performance of classification models. Specificity is the fraction of poorly-absorbed compounds correctly classified by the model and is inversely proportional to the number of false positives (poorly-absorbed compounds wrongly classified as highly-absorbed compounds). Specificity is defined as SP = TN/(TN + FP), where TN is the number of true negatives (poorly-absorbed compounds correctly classified as poorly-absorbed) and FP is the number of false positives. Sensitivity is the ratio of highly-absorbed compounds correctly classified by the model, and is inversely proportional to the number of false negatives. Sensitivity is defined as SE = TP/(TP + FN), where TP is the number of true positives (highly-absorbed compounds correctly classified as highly-absorbed) and FN is the number of false negatives (highly-absorbed compounds wrongly classified as poorly-absorbed compounds). The overall predictive performance of a model was measured by multiplying the specificity and sensitivity (SP x SE). This is an effective measure of a model's predictive performance as it takes into account the effect of unbalanced class distribution. In contrast, the overall accuracy measure, usually defined by the ratio of the number of correct predictions made by the model over the total number of (correct or wrong) predictions, does not take into account the effect of unbalanced class distributions or misclassification costs. To take into account misclassification costs in the models, the cost normalised misclassification index (CNMI) was calculated. CNMI can be calculated by **Equation 1** below.

$$CNMI = \frac{(\mathbf{FP} \times \mathbf{Cost_{FP}}) + (\mathbf{FN} \times \mathbf{Cost_{FN}})}{(\mathbf{Neg} \times \mathbf{Cost_{FP}}) + (\mathbf{Pos} \times \mathbf{Cost_{FN}})} \qquad \textbf{Eq. 1}$$

CostFP and CostFN are the misclassification costs assigned for false positives and false negatives and Neg and Pos define the total number of negative and positive observations, respectively. The CNMI value will be between zero and one, zero showing no misclassification errors and as the number increases towards one the number of misclassifications increases. For a more detailed explanation of **Equation 1**, see reference [28]

3. Results

A full list of molecular descriptors selected by each of the feature selection methods can be found in the supporting information (**Supporting information**). For GRD and GEN, as these

are not ranking feature selection methods the number of descriptors picked by the method will depend on the technique. GRD selected a total of 21 descriptors and GEN selected 64. **Tables 3** and **4** show the predictive performance measures from the classification trees using different sets of molecular descriptors from feature selection methods. In **Table 3** equal misclassification costs have been applied to false positive and false negatives for C&RT analysis, while in **Table 4** the ratio of misclassification costs is 4:1 for false positives: false negatives. In **Table 3** and **4** the best models are those that have the highest SE, SP and SP x SE measures and the lowest CNMI. These have been highlighted in **bold** for the training (t), parameter optimisation (po) and validation (v) sets. For the random forest feature selection method there was an option to apply misclassification costs. Therefore the descriptor sets selected by RF with equal (models 1 and 8) and higher misclassification costs applied to false positives (models 2 and 9) were used and also compared. All the C&RT decision trees from **Tables 3 and 4** can be found in the **Supporting Information**.

Table 3. The results of C&RT classification analysis using different feature selection methods with equal misclassification costs applied to the C&RT algorithm

Model	Feature selection Method	dataset	N	SP x SE	SE	SP	CNMI
1	RF	t	531	0.848	0.950	0.892	0.060
		po	107	0.709	0.930	0.762	0.103
		v	47	0.363	0.816	0.444	0.255
2*	RF (MC)	t	531	**0.884**	0.945	**0.935**	0.056
		po	107	**0.757**	0.884	**0.857**	0.121
		v	47	**0.453**	0.816	**0.556**	0.234
3	CS	t	531	0.777	0.963	0.806	0.064
		po	107	0.576	0.930	0.619	0.131
		v	47	0.187	0.842	0.222	0.277
4	IGR	t	531	0.800	**0.979**	0.817	0.049
		po	107	0.664	0.930	0.714	0.112
		v	47	0.398	**0.895**	0.444	**0.191**
5	GRD	t	531	0.803	0.970	0.828	0.055
		po	107	0.628	**0.942**	0.667	0.112
		v	47	0.351	0.789	0.444	0.277
6	GEN	t	531	0.839	0.975	0.860	**0.045**
		po	107	0.673	**0.942**	0.714	0.103
		v	47	0.398	**0.895**	0.444	**0.191**

7	C&RT	t	531	0.784	0.959	0.817	0.066
		po	105	0.694	**0.942**	0.737	**0.095**
		v	47	0.281	0.842	0.333	0.255

SE= Sensitivity, SP = Specificity; SP × SE = accuracy; CNMI = Cost normalised misclassification index, * misclassification costs applied to feature selection method

Comparing models built with equal misclassification costs (**Table 3**); the best overall model to choose would be model 2. This model has the highest SP x SE, plus the highest specificity values for the training, parameter optimisation and validation sets. However, this model does not achieve the highest sensitivity values, with SE = 0.945, 0.884 and 0.816 for the training, parameter optimisation and validation set respectively. All other models have better SE than model 2 for the three data subsets; apart from model 1, which has the same SE for the validation set, and model 5 (GRD), with a lower SE of 0.789. If the aim of the model was to achieve the best sensitivity then model 6, using genetic search feature selection, would be the best model to use as it achieved the best sensitivity for the parameter optimisation and the highest SE for the training set amongst the three selected models above, along with the lowest CNMI for the training set. Model 2 was able to classify correctly all the permanent ammonium-containing compounds used in the training and parameter optimisation set, and this reflected in the correct prediction of a permanent ammonium containing compounds in the validation set. The classification tree using the molecular descriptors from this model is shown in **Figure 3**.

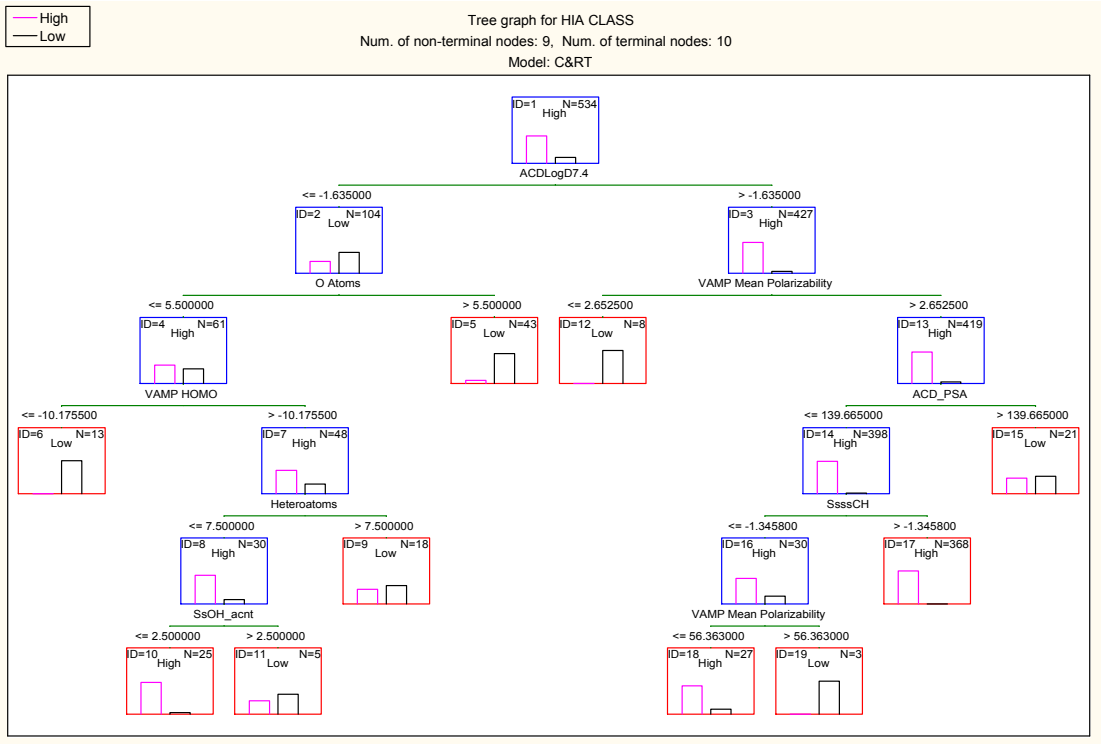


Figure 3. Tree graph for C&RT analysis using random forest predictor importance as feature selection method with equal misclassification costs applied to pre-processing C&RT (Model 2 in Table 3)

Table 4. The results of C&RT classification analysis using different feature selection methods with higher misclassification costs applied to false positives to the C&RT algorithm (misclassification cost ratio of FP: FN = 4:1)

Model	Feature selection Method	dataset	N	SP x SE	SE	SP	CNMI
8	RF	t	531	0.887	0.927	0.957	0.026
		po	107	0.725	0.895	0.810	0.068
		v	47	0.675	**0.868**	0.778	0.081
9*	RF (MC)	t	531	0.879	0.909	0.968	0.028
		po	107	**0.738**	0.860	**0.857**	**0.066**
		v	47	0.635	0.816	0.778	0.093
10	CS	t	531	0.838	0.906	0.925	0.037
		po	107	0.687	0.849	0.810	0.079
		v	47	0.544	0.816	0.667	0.118
11	IGR	t	531	0.853	0.934	0.914	0.033
		po	107	0.673	0.884	0.762	0.082

		v	47	0.544	0.816	0.667	0.118
12	GRD	t	528	0.892	**0.943**	0.946	0.025
		po	106	0.654	0.872	0.750	0.085
		v	47	**0.725**	0.816	**0.889**	**0.068**
13	GEN	t	531	0.885	0.895	**0.989**	0.027
		po	107	0.640	0.895	0.714	0.090
		v	47	0.614	0.789	0.778	0.099
14	C&RT	t	531	**0.911**	0.932	0.978	0.020
		po	107	0.726	**0.907**	0.800	**0.066**
		v	47	0.544	0.816	0.667	0.118

FP = False positive; FN = False negative; SE= Sensitivity, SP = Specificity; SP × SE = accuracy; CNMI = Cost normalised misclassification index, * misclassification costs applied to feature selection method

For **Table 4**, based on the SP x SE for the external validation set the best model is model 12 with a SP x SE value of 0.725 but this model also had one of the lowest SP x SE for the po set (0.654) which has a higher number of chemicals compared to the external validation set. In comparison, models 8 and 9 achieved higher SP x SE of 0.725 and 0.738 respectively, where the po set was not used for molecular descriptor selection and hence it was also an external set. From models 8 and 9, model 9 had a similar balance of high estimation of SP and SE compared to model 8 which was slightly worse at predicting poorly-absorbed compounds for the po set. What was interesting to note about model 9 was the feature selection method using predictor importance from random forest, which allowed misclassification costs to be applied at the feature selection level. Then the resulting C&RT model (with misclassification costs) achieved high prediction accuracy for the unseen validation set as well as training and parameter optimisation sets. The C&RT tree for model 9 is shown in **Figure 4**.

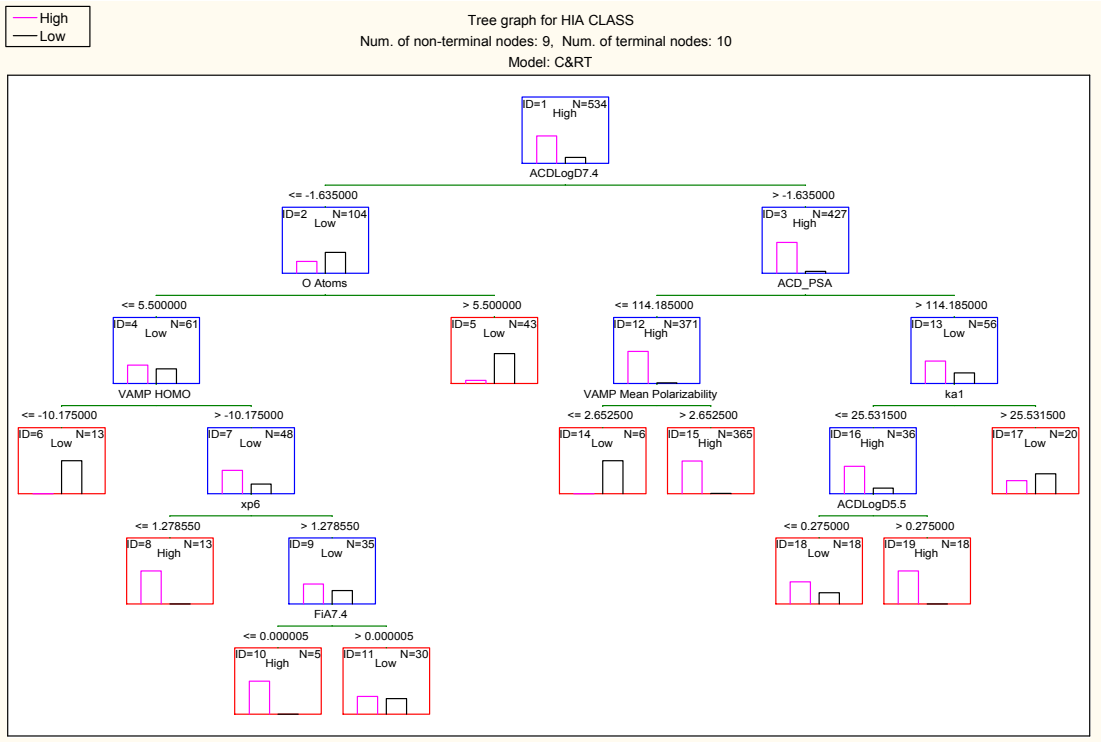


Figure 4. Tree graph for C&RT analysis using random forest predictor importance as feature selection method with higher misclassification costs applied to reduce false positives (model 9 in Table 4)

3.1 Interpretation of the selected models (models 2 and 9)

Both models 2 and 9 have been developed using the 20 most significant molecular descriptors selected by random forest analysis. Although the top 20 molecular descriptors were given as input to the C&RT analysis, not all of the molecular descriptors were used to build the decision trees. The first split variable in both models is ACDLogD7.4, the logarithm of the apparent distribution coefficient between octanol and water, and a measure of hydrophobicity at pH 7.4. This molecular descriptor along with logP is used in numerous publications for oral absorption modelling, and has been found to have a positive effect for transcellular absorption [43, 44]. For compounds to be split into the high absorption class, LogD7.4 has to be greater than -1.63 according to both models. For compounds with low logD7.4 ($\leq$-1.63), if they contain more than five oxygen atoms they are classed as poorly-absorbed in this terminal node according to both models. This molecular descriptor is linked to the number of hydrogen bond acceptors, highlighted in Lipinski's rule of five [45]; which states that a molecule will be highly likely to be poorly-absorbed if two or more of the following rules are

broken: if molecular weight >500 Da, sum of OH and NH hydrogen bond donors >5, calculated logP (C LogP) >5 and sum of N and O atoms as hydrogen bond acceptors >10. Examples of poorly-absorbed compounds classed in this node are ceftriaxone and raffinose.

In both models, the next important descriptor selected for the partitioning of compounds with low logD7.4 and less than six oxygen atoms is VAMP HOMO. This molecular descriptor is the energy of the highest occupied molecular orbital calculated by AM1 semi-empirical method using the VAMP programme in TSAR 3D software. The higher the value (>-10.18 in the split in the trees) indicates higher absorption classification. Compounds with low HOMO values are in the low absorption terminal nodes; and correlates with previous research [28]. The majority of compounds with low HOMO energy (<-10.18) according to this split contain a permanent quaternary ion such as pralidoxime and bethanechol, which are small polar molecules mainly related to the neurotransmitter acetylcholine, or compounds such as fosmidomycin and fosfomycin, which contain phosphorus atoms. Compounds with a higher HOMO energy are further split with different molecular descriptors in the two trees.

In **Figure 3** compounds with more than seven heteroatoms are classed as poorly-absorbed. This corresponds to Lipinski's rule of five, more precisely the number of hydrogen bond acceptors rule. In this node the majority of compounds are antibiotics such as meropenem and imipenem, which are both poorly-absorbed. There are also some misclassified antibiotics such as penicillin V and amoxicillin, which are highly-absorbed. However, both these compounds have been found to be substrates for the oligopeptide transporter, PEPT1 (SLC15A1), influx transporter in the small intestine [46]. The remaining 30 compounds are classed as highly-absorbed if they contain less than three OH groups (SsOH_Acnt).

In **Figure 4** however, compounds with low xp6 values are classed as highly-absorbed. The descriptor xp6 is the sixth order single path molecular connectivity index [9], which may be regarded as a size descriptor with some shape/connectivity elements. Examples of compounds in this node are of a small, polar often peptide like nature with no permanent charge and mainly natural or semi-synthetic compounds such as phenylalanine and captopril which may have the possibility to be absorbed using oligopeptide transporters (**Figure 4**, Node ID=8). The remaining 35 compounds are classed as poorly-absorbed if they have acidic groups with ionization fraction > 0.000005.

Highly-absorbed compounds with logD value greater than -1.63 are split differently in **Figures 3** and **4**. Despite this, the best molecular descriptors for splitting of these 427

compounds in both trees are the same, namely polarizability (VAMP mean polarizability) and polar surface area (ACD_PSA). In both trees, compounds with polarizability values ≤ 2.65 are poorly-absorbed. This molecular descriptor indicates the distortion of a compound's electron cloud by an external electric field [47]. Examples of compounds with ≤2.65 polarizability values (Node ID = 12 in **Figure 3** and 14 in **Figure 4**) are bephenium and vecuronium, both with low polarizability due to the permanent quaternary ion present in the molecules. Polar surface area (PSA) is a common molecular descriptor used in oral absorption models [20, 28]. PSA is the area of the Van der Waals surface that arises from oxygen and nitrogen atoms or hydrogen atoms bound to these atoms. In both trees compounds with high PSA are poorly-absorbed. In **Figure 3** a compound is poorly-absorbed if the PSA is greater than 139.67Å, which matches the literature threshold value where it was cited that a molecule will be poorly-absorbed (<10% FA) if the PSA is ≥140Å [43, 48]. In **Figure 4**, a threshold value of 114.19 Da has been used but these high PSA compounds have been partitioned further and those with smaller molecular size as indicated by ka1, and higher logD5.5 values than 0.275 are classed as highly-absorbed. An interesting feature can be observed in **Figure 3**, where for the compounds with PSA values ≤139.67 and low index for >CH- groups (SsssCH ≤ -1.35), if polarizability is too high (VAMP mean polarizability >56.363) then oral absorption will be poor. Examples of these drugs are two pro-drug ACE inhibitors moexipril diacid and fosinopril plus the cardiac glycoside cymarin.

3.2 Chemical space and repeating misclassifications in models

There were a few compounds that were continually misclassified by most models. Within the validation set the compounds misclassified by all models was lovastatin, while frovatriptan was misclassified by the majority of models. These compounds are poorly-absorbed, but the models misclassified them as highly-absorbed. Lovastatin is a naturally occurring product used to reduce cholesterol; this compound has poor solubility issues in aqueous medium [49], plus it has been identified as heavily undergoing gut metabolism both of which could account for the misclassification [50]. In addition, this compound has been identified as a potential substrate and inhibitor of the efflux transporter P-gp [51]. Frovatriptan, according to the Varma et al (2010), has a fraction escaping gut metabolism of 69% meaning potentially, 30% could be metabolised by the gut, specifically UDP-glucuronosyltransferases (UGT's) in the gut due to their substrate specificity of the indole group present in frovatriptan and the similarity of this compound to serotonin, a UGT substrate. However there is no direct evidence of this in the literature however this could explain the misclassification by our models [52, 53].

4. Discussion

Thousands of molecular descriptors can be calculated to represent molecular features or properties of the compounds. The use of feature selection to reduce the number of molecular descriptors is a common practice in QSAR as a part of pre-processing or embedded methods. Feature selection increases interpretability by reducing the number of molecular descriptors, it reduces overfitting associated with noisy or redundant descriptors and often improves predictability of resulting classification models.

In this paper we used various filter feature selection methods for data pre-processing, to pick significant descriptors related to intestinal absorption. These descriptor sets were used as input for C&RT analysis, which has an embedded feature selection method, to classify compounds into high or low absorption in a biased dataset. The application of higher misclassification costs for false positives to the C&RT analysis was also investigated to overcome the problem of biased datasets (which contain many more highly-absorbed compounds than poorly-absorbed compounds) and to see if models with greater prediction accuracy could be achieved.

The feature selection methods used in this work were predictor importance using random forest (RF), chi square (CS), information gain ratio (IGR), greedy search (GRD) and genetic search (GEN). The feature selection methods were compared based on the predicted ability of the C&RT algorithm. There were certain expectations of the feature selection methods based on how they work and their advantages and disadvantages. To begin, it was expected that the combination of a pre-processing feature selection method and C&RT, which has an embedded feature selection, to have higher prediction accuracy when compared to using C&RT with no pre-processing feature selection method. This was on the basis that when C&RT splits compounds, further down the tree there are fewer compounds in the deeper nodes, therefore less statistical support for an effective selection of the best descriptor especially when there are a larger number of molecular descriptors to choose from. Therefore as a result the C&RT algorithm could pick descriptors that may be less relevant to molecular descriptors higher up in the tree. However, C&RT is a successful technique in its own right with an embedded feature selection function which is used in model development for the prediction of oral absorption [28, 54]. The benefits of using C&RT are that it can cope with noisy data (to some extent) of moderately sized biased datasets [11] and produces models (decision

trees) that in principle can be easily interpreted. In addition it is less time consuming than pre-processing the molecular descriptors first.

The expectations of the feature selection methods themselves can be considered and compared to the obtained results in this work. The benefits of simple univariate filter techniques such as CS and IGR are that they are simple and fast to compute; however they fail to take into account molecular descriptor interactions [18, 55]. This is in contrast to GRD and GEN, which take molecular descriptor interactions into account but are more computationally expensive. In a comparison of GRD and GEN, due to the way these feature selection methods work, GEN should achieve higher accuracy, as it performs a global search in the molecular descriptor space, whilst GRD performs a local search in the molecular descriptor space. Using the predictor importance in the random forest method is computationally expensive; however, there is the added advantage that misclassification costs can be applied using the software as well as being applied for the C&RT analysis. Finally, based on previous research, the application of higher misclassification costs to false positives will produce models with increased overall accuracy and reduced false positive misclassifications, therefore overcoming the problem of biased datasets compared with equal misclassification costs.

Overall, one of the best feature selection methods according to the models produced in this work was predictor importance using random forest. This was expected for this method, as it was possible to apply higher misclassification costs to the feature selection technique itself as well as applied to the C&RT analysis. Even when misclassification costs were not applied to predictor importance, the produced models still had higher overall accuracies over most models. This is down to the ensemble nature of this method, which is known to perform better than single tree analysis [56]. In comparison with C&RT where no pre-processing feature selection was utilised, the predictor importance feature selection method had higher overall accuracy for the validation set in all cases. The high classification accuracy on the training set but low prediction accuracy on the validation set could indicate overfitting of the models produced by C&RT. Models produced by other pre-processing feature selection techniques were better compared with models produced by C&RT with no pre-processing feature selection on the validation set, except for the models produced by CS feature selection. In the majority of the cases, using C&RT alone gave better prediction accuracy for the parameter optimisation set compared with IGR, GRD and GEN; however these latter methods had better overall prediction accuracy for the validation set. This shows that C&RT without pre-processing can cope with redundant and meaningless molecular descriptors, however is prone

to overfitting (even with pruning of the trees) and can lack predictive accuracy for the prediction of the validation set.

Comparing the expectations set out initially, it was found that comparing univariate methods such as CS and IGR with those that take into account molecular descriptor interactions (GEN and GRD) there is no clear pattern in the difference between their results. However, overall, when equal misclassification costs were applied to the C&RT analysis, GEN as expected had better or comparable predictor performance than CS and IGR. On the other hand, GRD had comparable performance with CS and weaker compared with IGR. When misclassification costs were applied to C&RT, GEN and GRD models were better than CS and IGR for the training and validation set, again it is difficult to state which method is better overall. This effect is also seen in the next example when comparing GEN and GRD feature selection methods based on the predictive accuracy of the C&RT analysis. GEN performs better than GRD when equal misclassification costs were used; this matches the predictions previously made. This is in some agreement with work using numerical regression analysis [57]. However upon applying higher misclassification costs the molecular descriptors pre-processed by the GRD model outperformed the GEN model. This could be due to the correlation-based feature selection subset evaluator used by the GEN method not being suitable for use with C&RT and misclassification costs, and potentially highlight overfitting by the GEN based model. The effect of applying higher misclassification costs to either false positives or false negatives has been investigated in previous research [28]. In this work the application of higher misclassification costs to false positives resulted in better overall accuracy and specificity as expected in the majority of cases.

In this work we have shown that for most models using pre-processing feature selection does appear to improve classification accuracy compared to the control (C&RT using all molecular descriptors) based on prediction accuracy. This agrees with work carried out by Xue and co-workers[16] who considered three different datasets including prediction of oral absorption. They used recursive feature elimination for feature selection and SVM to classify compounds. They compared the results with and without the feature selection method and found that for oral absorption improved accuracy was obtained when the feature selection method was used. For one of the datasets, feature selection gave comparable predictive ability, which with a smaller descriptor subset will increase the interpretability of resulting models [16]. However, a study by Suenderhauf[11] carried out regression and classification for oral absorption using a variety of techniques including C&RT, Support Vector Machine

(SVM), and chi-squared automatic interactor detector (CHAID); and using these techniques they compared the feature selection methods of best first feature selection (BFS) using a greedy hill-climbing algorithm, linear correlation analysis and decision tree splitting criteria. Suenderhauf utilised the decision trees to pick smaller subsets of molecular descriptors used in the work as input for the model development, this is similar to our idea of a two stage pre-processing feature selection. The best model was produced by CHAID using the entire set of original molecular descriptors, which contradicts our results that pre-processing feature selection gives better accuracy. However, it is interesting to note that, out of the feature selection methods that they used, the decision tree splitting criteria gave the best results. This research also showed that SVM had poor performance when feature selection methods were utilised, which could indicate that there are some feature selection methods that work better with certain techniques such as SVM [11].

Although it is difficult to directly compare the different feature selection techniques that we used with the literature, the molecular descriptor subsets can be compared. Firstly it is interesting to compare in this work the molecular descriptors selected by the pre-processing feature selection methods (**Supporting information**). The top molecular descriptors picked by the feature selection methods can be found in **Table 5**. This table shows the top molecular descriptors that were picked by three or more feature selection methods. The molecular descriptors selected by the various feature selection methods were used as input for C&RT analysis, which in turn further selected a smaller subset of molecular descriptors to build decision trees. The top descriptors picked by firstly the pre-processing method and then by C&RT analysis are shown in **Table 6**. **Table 6** also indicates the number of times a molecular descriptor was picked by C&RT with or without pre-processing feature selection. The individual descriptors picked by the C&RT models can be found in the **supporting information**.

Table 5. Molecular descriptors selected by three or more pre-processing feature selection methods listed in Table 2

Descriptor	Feature selection method	Description
ACDLogD7.4	RF, RF (MC), CS, IGR, GRD, GEN	Apparent distribution coefficient at pH 7.4 calculated by ACD
ACDLogD10	RF, CS, IGR, GRD, GEN	Apparent distribution coefficient at pH 10 calculated by ACD
ACDLogD5.5	RF, RF (MC), CS, GRD, GEN	Apparent distribution coefficient at pH 5.5 calculated by ACD
SHHBd	CS, IGR, GRD, GEN	Sum of the hydrogen atom level E-state values for all hydrogen atoms bonded to donating atoms
O Atoms	RF, RF (MC), CS, GRD	Number of oxygen atoms in whole molecule
ACD_PSA	RF, RF (MC), CS, GRD	Polar surface area

numHBa	RF, RF (MC), CS, GEN	Number of Hydrogen bond acceptors
SsOH_acnt	RF, RF (MC), CS, GEN	Counts of atom-type E-state for hydroxyl groups
VAMP Heat of Formation	RF, RF (MC), GRD, GEN	Enthalpy required to form 1 mole of compound at 298K calculated by VAMP
ACD_LogP	CS, GRD, GEN	Octanol/water partition coefficient calculated by ACD
ACDLogD6.5	RF, CS, GRD	Apparent distribution coefficient at pH 6.5 calculated by ACD
Heteroatoms	RF (MC), CS, GEN	Number of atoms that are not carbon or hydrogen e.g nitrogen, oxygen
ka1	RF, RF (MC), GEN	First order kappa alpha shape index
numHBd	RF, CS, GEN	Number of hydrogen bond donors
SdsssP	IGR, GRD, GEN	Sum of atom-type E-state for phosphorous atoms with 3 single and one double bond
Sum of E-State indices	RF, IGR, GEN	Sum of the E-State values for all the atoms in molecule
VAMP HOMO	RF (MC), GRD, GEN	Energy of the highest occupied molecular orbital calculated by VAMP
VAMP LUMO	RF, GRD, GEN	Energy of the lowest occupied molecular orbital calculated by VAMP

Table 6. The top molecular descriptors selected by C&RT

		Times used by C&RT	
Type of descriptor	Descriptor	Pre-processing	No Pre-processing
Hydrogen bonding	ACD_PSA	8	1
	O Atoms	8	1
	SHHBd	8[a]	
Lipophilicity	ACDLogD7.4	10	1
	ACD_LogP	6	
	ACDLogD6.5	3	1
Polarity/ Polarization	VAMP LUMO	4[a]	2
	N+	5[a]	
	VAMP Mean Polarizability	5[a]	
Size/Shape	VAMP totl Energy	5[a]	
	ka1	3[a]	
	SsssCH	3	1

[a]Occurred more than once in a single tree model.

The top molecular descriptor as picked by the majority of feature selection methods was the same as the top molecular descriptor then picked by the resulting C&RT analysis (ACDLogD7.4). Other studies have identified lipophilicity descriptors, in particular logD7.4 as well as logD5.5, 6.5 and logP, as important for intestinal absorption as picked by various feature selection techniques [11, 58, 59]. The next most frequently picked molecular descriptors are those relating to hydrogen bonding, in particular polar surface area. Polar surface area is a molecular descriptor commonly used in oral absorption models and it has a negative correlation with intestinal absorption [43, 48]. This descriptor was also utilised in other studies that focussed on feature selection techniques as well as oral absorption modelling [11, 20, 59]. The other top hydrogen bonding descriptors highly ranked are the number of oxygen atoms and SHHBd, which is related to the number hydrogen bond donors in a molecule. Both these descriptors were picked by the feature selection models and utilised in the C&RT analysis high up near the tree root indicating the importance of these descriptors. Descriptors relating to hydrogen bonding capacity are important in oral absorption modelling and are used in the

widely accepted filter, Lipinski's rule of five [45]. Overall the top descriptors picked by the feature selection methods and then utilised by C&RT are very similar. Also, the majority of molecular descriptors used by C&RT without any pre-processing feature selection match those picked by the pre-processing feature selection methods, with a few exceptions. The top descriptors in **Table 6** are in line with the literature where among these molecular descriptors related to absorption are those that describe lipophilicity, molecular size/shape, polar surface area, hydrogen bonding, and similar parameters.

5. Conclusion

Feature selection is important in its many forms as a way to increase interpretability and predictability but reduce over-fitting of QSAR models. This work has shown that pre-processing filter feature selection methods can greatly improve QSAR models using C&RT analysis. C&RT can be used as an embedded feature selection method, however it can be inadequate since further down the tree there are fewer compounds available for descriptor selection and therefore descriptors may be selected which are not optimal. Here we have used several pre-processing feature selection methods prior to C&RT and have produced more accurate QSAR models for the estimation of oral absorption class as shown by the external sets of compounds. However, examination of the literature reveals that different feature selection methods utilised with different classification methods should be tried and evaluated for one dataset. Similar molecular descriptors were picked by the different feature selection methods; and those descriptors relate to lipophilicity, hydrogen bonding, polarity, size and shape. Higher misclassification costs applied to reduce false positives yielded models with better overall predictability of highly and poorly-absorbed compounds. The use of filter pre-processing feature selection methods and misclassification costs produce models with better interpretability and predictability that overcome the problem of a biased dataset with many more highly-absorbed compounds than poorly-absorbed compounds and shows the importance of feature selection in QSAR model development.

Supporting Information

The supporting information contains a list of the 47 compounds in the external validation set and their HIA% values **(S1)**, a list of molecular descriptors picked by the feature selection methods **(S2)**, full list of all the molecular descriptors picked by C&RT analysis **(S3)** and finally all the C&RT decision trees produced from this work **(S4)**. This information is available free of charge via the Internet at http://pubs.acs.org.

References

1. DiMasi, J. A.; Hansen, R. W.; Grabowski, H. G., The price of innovation: new estimates of drug development costs. *J. Health. Econ.* **2003**, *22*, 151-185.

2. DiMasi, J. A., Risks in new drug development: Approval success rates for investigational drugs. *Clin. Pharmacol Ther.* **2001**, *69*, 297-307.

3. Bunnage, M. E., Getting pharmaceutical R&D back on target. *Nat. Chem Bio.* **2011**, *7*, 335-339.

4. Kola, I.; Landis, J., Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discovery.* **2004**, *3*, 711-715.

5. Ashford, M., Part 4: Biopharmaceutical principles of drug delivery. In *Aulton's Pharmaceutics, The design and manufacture of medicines*, third edition.; Aulton, M. E., Ed. Churchill Livingstone Elsevier: Philadelphia, **2007**; pp 265-324.

6. Hou, T. J.; Li, Y. Y.; Zhang, W.; Wang, J. M., Recent Developments of In Silico Predictions of Intestinal Absorption and Oral Bioavailability. *Comb. Chem. High Throughput Screening.* **2009**, *12*, 497-506.

7. van de Waterbeemd, H.; Gifford, E. ADMET in silico modelling: towards prediction paradise? *Nat. Rev. Drug Discovery*. **2003**, *2*, 192-204.

8. Todeschini, R.; Consonni, V., *Handbook of Molecular Descriptors*. first edition.; Wiley-VCH: Verlag, 2000.

9. Hall, L. H.; Kier, L. B., The Molecular Connectivity Chi Indices and Kappa Shape Indices in Structure-Property Modeling. In *Reviews in Computational Chemistry*, Boyd, D.; Lipkowitz, K., Eds. VCH: New York, 1991; pp 384–385.

10. Ertl, P.; Rohde, B.; Selzer, P., Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.* **2000**, *43*, 3714-3717.

11. Suenderhauf, C.; Hammann, F.; Maunz, A.; Helma, C.; Huwyler, J., Combinatorial QSAR Modeling of Human Intestinal Absorption. *Mol. Pharmaceutics.* **2011**, *8*, 213-224.

12. Wong, W. W. L.; Burkowski, F. J., Using Kernel Alignment to Select Features of Molecular Descriptors in a QSAR Study. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2011**, *8*, 1373-1384.

13. Ghafourian, T.; Cronin, M. T. D., The impact of variable selection on the modelling of oestrogenicity. *SAR QSAR Environ. Res.* **2005**, *16*, 171-190.

14. Liu, Y., A comparative study on feature selection methods for drug discovery. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1823-1828.

15. Dudek, A. Z.; Arodz, T.; Galvez, J., Computational methods in developing quantitative structure-activity relationships (QSAR): A review. *Comb. Chem. High Throughput Screening.* **2006**, *9*, 213-228.

16. Xue, Y.; Li, Z. R.; Yap, C. W.; Sun, L. Z.; Chen, X.; Chen, Y. Z., Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1630-1638.

17. Goodarzi, M.; Dejaegher, B.; Vander Heyden, Y., Feature Selection Methods in QSAR Studies. *J. AOAC. Int.* **2012**, *95*, 636-651.

18. Saeys, Y.; Inza, I.; Larranaga, P., A review of feature selection techniques in bioinformatics. *Bioinformatics.* **2007**, *23*, 2507-2517.

19. Kohavi, R.; John, G. H., Wrappers for feature subset selection. *Artif. Intell.* **1997**, *97*, 273-324.

20. Wegner, J. K.; Frohlich, H.; Zell, A., Feature selection for descriptor based classification models. 2. Human intestinal absorption (HIA). *J. Chem. Inf. Comput. Sci.* **2004**, 44, 931-939.

21. Hou, T. J.; Wang, J. M.; Zhang, W.; Xu, X. J., ADME evaluation in drug discovery. 7. Prediction of oral absorption by correlation and classification. *J. Chem. Inf. Model.* **2007**, *47*, 208-218.

22. Hou, T. J.; Wang, J. M.; Li, Y. Y., ADME evaluation in drug discovery. 8. The prediction of human intestinal absorption by a support vector machine. *J. Chem. Inf. Model.* **2007**, *47*, 2408-2415.

23. Niwa, T., Using general regression and probabilistic neural networks to predict human intestinal absorption with topological descriptors derived from two-dimensional chemical structures. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 113-119.

24. Wessel, M. D.; Jurs, P. C.; Tolan, J. W.; Muskal, S. M., Prediction of human intestinal absorption of drug compounds from molecular structure. *J. Chem. Inf. Comput. Sci.* **1998,** *38*, 726-735.

25. Ku, M. S., Use of the biopharmaceutical classification system in early drug development. *AAPS J.* **2008**, *10*, 208-212.

26. Wu, C. Y.; Benet, L. Z., Predicting drug disposition via application of BCS: Transport/absorption/elimination interplay and development of a biopharmaceutics drug disposition classification system. *Pharm. Res.* **2005,** *22*, 11-23.

27. Ghafourian, T.; Newby, D.; Freitas, A. A., The impact of training set data distributions for modelling of passive intestinal absorption. *Int. J. Pharm.* **2012,** 436, 711-720.

28. Newby, D.; Freitas, A. A.; Ghafourian, T., Coping with Unbalanced Class Data Sets in Oral Absorption Models. *J. Chem. Inf. Model.* **2013,** *53*, 461-474.

29. Breiman, L.; Friedman, J.; Stone, C. J.; Olshen, R. A. *Classification and Regression Trees.* First edition; Chapman and Hall/CRC: Boca Raton, **1984**.

30. Tan, P. N.; Steinbach, M.; Kumar, V., *Introduction to Data Mining*. first edition.; Pearson International Edition: Boston, **2006**.

31. Wold, S.; Berglund, A.; Kettaneh, N., New and old trends in chemometrics. How to deal with the increasing data volumes in R&D&P (research, development and production) - with examples from pharmaceutical research and process modeling. *J. Chemom.* **2002**, *16*, 377-386.

32. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H., The WEKA Data Mining Software: An Update. *SIGKDD Explorations.* **2009**, 11, 10-18.

33. Breiman, L., Random forests. *Mach. Learn.* **2001**, *45*, 5-32.

34. Liaw, A.; Wiener, M., Classification and Regression by randomForest. *R News* **2002,** 2/3, 18-22.

35. Liu, H.; Setiono, R., Chi2: Feature selection and discretization of numeric attributes. In *Seventh International Conference on Tools with Artificial Intelligence*, Herndon, Virginia, Nov 5-8, 1995; Vassilopoulos, J. F., Ed.; IEEE Computer Society Press, Washington; pp 388-391, 1995

36. Martinez, M. N.; Amidon, G. L., A mechanistic approach to understanding the factors affecting drug absorption: A review of fundamentals. *J. Clin Pharmacol.* **2002**, *42*, 620-643.

37. Quinlan, J. R., *Discovering rules from large collections of examples: A case study*. first edition.; Edinburgh University Press.: Edinburgh, **1979**.

38. Quinlan, J. R., *C4.5: programs for machine learning*. first edition.; Morgan Kaufmann Publishers Inc: San Francisco, 1993.

39. Kittler, J., Feature set search algorithms. In *Pattern Recognition and Signal Processing*, Paris, France, June 25th -4th July, 1978, Chen, C. H., Ed. Sijthoff and Noordhoff,: The Netherlands, pp 41-60, **1978**.

40. Shah, S. C.; Kusiak, A., Data mining and genetic algorithm based gene/SNP selection. *Artif. Intell. Med.* **2004**, *31*, 183-196.

41. Holland, J. H., *Adaptation in Natural and Artificial Systems*. University of Michigan Press (re-issued by MIT Press 1992): Michigan, **1975**.

42. Goldberg, D. E., *Genetic algorithms in search, optimization, and machine learning*. first edition.; Addison-wesley Longman Publishing: Boston, 1989.

43. Clark, D. E., Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 1. Prediction of intestinal absorption. *J. Pharm. Sci.* **1999,** *88*, 807-814.

44. Hou, T. J.; Wang, J. M.; Zhang, W.; Wang, W.; Xu, X., Recent advances in computational prediction of drug absorption and permeability in drug discovery. *Curr. Med. Chem.* **2006,** *13*, 2653-2667.

45. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J., Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997,** *23*, 3-25.

46. Brandsch, M.; Knutter, I.; Bosse-Doenecke, E., Pharmaceutical and pharmacological importance of peptide transporters. *J. Pharm. Pharmacol.* **2008**, 60, 543-585.

47. Wang, J. M.; Xie, X. Q.; Hou, T. J.; Xu, X. J., Fast approaches for molecular polarizability calculations. *J. Phys. Chem A.* **2007,** *111*, 4443-4448.

48. Palm, K.; Stenberg, P.; Luthman, K.; Artursson, P., Polar molecular surface properties predict the intestinal absorption of drugs in humans. *Pharm. Res.* **1997,** *14*, 568-571.

49. Serajuddin, A. T. M.; Ranadive, S. A.; Mahoney, E. M., Relative lipophilicities, solubilities, and structure-pharmacological considerations of 3-hydroxy-3-methylglutaryl-coenzyme A (HMG-CoA) reductase inhibitors pravastatin, lovastatin, mevastatin, and simvastatin. *J. Pharm Sci.* **1991,** *80*, 830-834.

50. Jacobsen, W.; Kirchner, G.; Hallensleben, K.; Mancinelli, L.; Deters, M.; Hackbarth, I.; Baner, K.; Benet, L. Z.; Sewing, K. F.; Christians, U., Small intestinal metabolism of the 3-hydroxy-3-methylglutary-coenzyme A reductase inhibitor lovastatin and comparison with pravastatin. *J. Pharmacol. Exp. Ther.* **1999,** *291*, 131-139.

51. Wang, E. J.; Casciano, C. N.; Clement, R. P.; Johnson, W. W., HMG-CoA reductase inhibitors (statins) characterized as direct inhibitors of P-glycoprotein. *Pharm. Res.* **2001,** *18*, 800-806.

52. Varma, M. V. S.; Obach, R. S.; Rotter, C.; Miller, H. R.; Chang, G.; Steyn, S. J.; El-Kattan, A.; Troutman, M. D., Physicochemical Space for Optimum Oral Bioavailability: Contribution of Human Intestinal Absorption and First-Pass Elimination. *J. Med Chem.* **2010,** *53*, 1098-1108.

53. Krishnaswamy, S.; Duan, S. X.; Von Moltke, L. L.; Greenblatt, D. J.; Court, M. H., Validation of serotonin (5-hydroxtryptamine) as an in vitro substrate probe for human UDP-glucuronosyltransferase (UGT) 1A6. *Drug Metab. Dispos.* **2003**, *31*, 133-139.

54. Deconinck, E.; Hancock, T.; Coomans, D.; Massart, D. L.; Vander Heyden, Y., Classification of drugs in absorption classes using the classification and regression trees (CART) methodology. *J. Pharm. Biomed. Anal.* **2005**, *39*, 91-103.

55. Guyon, I.; Elisseeff, A., An Introduction to Variable and Feature Selection. *JMLR,* **2003,** *3*, 1157-1182.

56. Dietterich, T. G., Ensemble methods in machine learning. In *First international workshop*, *Multiple classifier systems, Lecture Notes in Computer Science*, Cagliari, Italy, June 21–23, 2000; Kittler, J., Roli, F., Eds.; Springer Berlin Heidelberg, Berlin; pp 1-15.

57. Xu, L.; Zhang, W.-J., Comparison of different methods for variable selection. *Anal. Chim. Acta.* **2001,** *446*, 475-481.

58. Agatonovic-Kustrin, S.; Beresford, R.; Yusof, A. P. M., Theoretically-derived molecular descriptors important in human intestinal absorption. *J. Pharm. Biomed. Anal.* **2001,** *25*, 227-237.

59. Winiwarter, S.; Bonham, N. M.; Ax, F.; Hallberg, A.; Lennernas, H.; Karlen, A., Correlation of human jejunal permeability (in vivo) of drugs with experimentally and theoretically derived parameters. A multivariate data analysis approach. *J. Med. Chem.* **1998**, *41*, 4939-4949.

For Table of Contents Use Only

Pre-processing feature selection for improved C&RT models for oral absorption

Danielle Newby, Alex. A. Freitas, Taravat Ghafourian*